**AN INTERNSHIP REPORT ON**

# STUDENT'S MARKS ANALYSIS

***Submitted by***


| | |
|---|---|
| **KAMIREDDY VINISH REDDY** | **(113219031068)** |
| **KIRTHICK RAJ S** | **(113219031074)** |
| **RAGHUL SAYEE K A D** | **(113219031115)** |
| **AKSHAYA BAALAJI S** | **(113219031006)** |


***In partial fulfillment for the award of the degree***
***Of***

**BACHELOR OF ENGINEERING**

**IN**

**COMPUTER SCIENCE AND ENGINEERING**



**VELAMMAL ENGINEERING COLLEGE, CHENNAI-66.**

(An Autonomous Institution)


**2021-2022**

# VELAMMAL ENGINEERING COLLEGE, CHENNAI-66

## BONAFIDE CERTIFICATE

Certified that this internship report **"STUDENTS' MARKS ANALYSIS"** is the bonafide work of **KAMIREDDY VINISH REDDY (113219031068)**, **KIRTHICK RAJ S (113219031074)**, **RAGHUL SAYEE K A D (113219031115)**, **AKSHAYA BAALAJI S (113219031006)** carried out at **MIT SQUARE** during 01.12.2021 to 31.01.2022.


**Dr. B. MURUGESWARI**
PROFESSOR & HEAD
Department of Computer science & Engineering
Velammal Engineering College
Chennai –600 066

**Dr. P. S. SMITHA**
ASSOCIATE PROFESSOR
Faculty Coordinator
Department of Computer science Engineering
Velammal Engineering College
Chennai – 600 066

# MIT SQUARE

**AN ISO 9001:2015 INTERNATIONAL CERTIFIED COMPANY**

**UK | USA | CANADA | SINGAPORE | MALAYSIA | AUSTRALIA | DUBAI | EUROPE**

## INTERNSHIP IN DATA SCIENCE

*This certifies that*

## VINISH REDDY KAMIREDDY

student from Velammal Engineering College has successfully completed two months of internship in the area of Data Science with our company, MIT Square Services Private Limited, from December 2021 to January 2022.

KNOWLEDGE IS POWER

100% CERTIFIED

S. Mithileysh

**DR MITHILEYSH SATHIYANARAYANAN**

FOUNDER | CEO | SCIENTIST

LONDON

WWW.MITSQUARE.COM

**CERTIFICATE#202201DS17**

# MIT SQUARE

**AN ISO 9001:2015 INTERNATIONAL CERTIFIED COMPANY**

**UK | USA | CANADA | SINGAPORE | MALAYSIA | AUSTRALIA | DUBAI | EUROPE**

## INTERNSHIP IN DATA SCIENCE

*This certifies that*

## KIRTHICK RAJ S

student from Velammal Engineering College has successfully completed two months of internship in the area of Data Science with our company, MIT Square Services Private Limited, from December 2021 to January 2022.

KNOWLEDGE IS POWER

100% CERTIFIED

**DR MITHILEYSH SATHIYANARAYANAN**

FOUNDER | CEO | SCIENTIST

LONDON

**WWW.MITSQUARE.COM**

**CERTIFICATE#202201DS15**

# MIT SQUARE

**AN ISO 9001:2015 INTERNATIONAL CERTIFIED COMPANY**

**UK | USA | CANADA | SINGAPORE | MALAYSIA | AUSTRALIA | DUBAI | EUROPE**

## INTERNSHIP IN DATA SCIENCE

*This certifies that*

## RAGHUL SAYEE

student from Velammal Engineering College has successfully completed two months of internship in the area of Data Science with our company, MIT Square Services Private Limited, from December 2021 to January 2022.

S Mithileysh

**DR MITHILEYSH SATHIYANARAYANAN**

**FOUNDER | CEO | SCIENTIST**

**LONDON**

**WWW.MITSQUARE.COM**

**CERTIFICATE#202201DS08**

# MIT SQUARE

**AN ISO 9001:2015 INTERNATIONAL CERTIFIED COMPANY**

**UK | USA | CANADA | SINGAPORE | MALAYSIA | AUSTRALIA | DUBAI | EUROPE**

## INTERNSHIP IN DATA SCIENCE

*This certifies that*

## AKSHAYA BAALAJI S

student from Velammal Engineering College has successfully completed two months of internship in the area of Data Science with our company, MIT Square Services Private Limited, from December 2021 to January 2022.


S. Mithileysh

**DR MITHILEYSH SATHIYANARAYANAN**

FOUNDER | CEO | SCIENTIST

LONDON

WWW.MITSQUARE.COM

**CERTIFICATE#202201DS19**

# CERTIFICATE OF EVALUATION

COLLEGE NAME : VELAMMAL ENGINEERING COLLEGE BRANCH

: COMPUTER SCIENCE AND ENGINEERING

SEMESTER : VI

| Sl. No | Name of the student who has done the Internship | Title of the Internship | Name of faculty coordinator with designation |
|---|---|---|---|
| 1 | KAMIREDDY VINISH REDDY | STUDENT'S MARKS ANALYSIS | Dr. P. S. SMITHA |
| 2 | KIRTHICK RAJ S | | |
| 3 | RAGHUL SAYEE K A D | | |
| 4 | AKSHAYA BAALAJI S | | |

This report of internship work submitted by the above students in partial fulfillment for the award of Bachelor of Computer Science and Engineering Degree in Velammal Engineering College was evaluated and confirmed to be reports of the work done by the above student and then assessed.

Submitted for Internal Evaluation held on.....

**Examiner 1** **Examiner 2** **Examiner** 3

## TABLE OF CONTENTS

## ABSTRACT:

EduTech (a combination of "education" and "technology") refers to hardware and software designed to enhance teacher-led learning in classrooms and improve students' education outcomes. The goal of EduTech is to improve student outcomes, enhance individualized education, and reduce the teaching burden on instructors. It encompasses several domains including learning theory, computer-based training, online learning and machine learning. EdTech is still in the early stages of its development, but it shows promise as a method of customizing a curriculum for a student's ability level by introducing and reinforcing new content at a pace the student can handle. The programming language is used to predict the student marks using machine learning is Python. In this paper we propose a Machine Learning (ML) approach that will be trained from the available data and gain intelligence and then uses the acquired knowledge for an accurate prediction.

This data set consists of the marks secured by the students in various subjects. To understand the influence of the parent's background, test preparation etc... on students' performance. This data set includes scores from three exams and a variety of personal, social, and economic factors that have interaction effects upon them.

Predictive analytics in education is all about knowing the mindset and needs of the students. It helps to make conclusions about the things that might happen in the future. With the class tests and half-yearly results, it could be understood which students are going to perform well in the exam and which students will have a tough time.

## LIST OF FIGURES:



## LIST OF TABLES:

## LIST OF SYMBOLS AND ABBREVIATION

| S.NO | SYMBOLS | ABBREVIATION |
|---|---|---|
| 1 | np | numpy |
| 2 | pd | pandas |
| 3 | mp | mathplot |
| 4 | df | dataframe |
| 5 | sb | Seaborn |
| 6 | mse | Mean Squared Error |
| 7 | rmse | Root Mean Squared Error |
| 8 | lin_reg | Linear Regression |
| 9 | r2 | Performance of regression |
| 10 | sc | Standard scalar |
| 11 | svr | Support Vector Regression |

# CHAPTER 1

# INTRODUCTION AND COMPANY PROFILE:

## 1.1 EDU TECH:

- Machine learning has the capability of better content and curriculum organization and management. It helps to bifurcate the work accordingly and understand the potential of everyone. This helps to analyze what work is best suited for the teacher and what works for the student.

- It also has the potential to make educators more efficient by completing tasks such as *classroom management, scheduling,* etc. Thus, the educators are free to focus on tasks that cannot be achieved by AI, and that require a human touch.

- Predictive analytics in education is all about knowing the mindset and needs of the students. It helps to make conclusions about the things that might happen in the future. With the class tests and half-yearly results, it
- could be understood which students are going to perform well in the exam and which students will have a tough time.

- Machine Learning makes it a lot easier for schools and colleges to better organize and manage their content and curriculum. It further helps to understand the potential of everyone in the system and then allocates the work accordingly. Needless to say, it helps improve the efficiency of the education system.

## 1.2 COMPANY PROFILE

MIT Square is a premier product development company headquartered in Bangalore, India and has a presence in Southampton, UK. MIT stands for "Management and Innovation for Transformation" with a tagline "We transform your life". MIT Square is an International Organization for Standardization, ISO 9001:2015, Certified Company. We at MIT Square, are experts in designing and developing innovative products, building start-ups, and understanding the need of the enterprises for their business growth. We offer product design, product development, product manufacturing and patent filing services.

MIT Square excels in the design, development, manufacturing and supplying of consumer products, industrial and IoT devices, education platforms, hospitality products, and healthcare technology. We offer turnkey, tooling and OEM/ODM services. From individuals, start-ups, small and medium-sized companies to international corporations, MIT Square is here to support you in all your product design & product development needs and pave the way to transform your life by turning your ideas into reality. MIT Square offers you an unparalleled equation of value, cost and on time delivery by having our highly qualified product design-development, supply chain and product manufacturing specialists team in UK, USA, Asia and Middle East. We have strong links with manufacturing units in India. From discovery to delivery, MIT Square is one stop solution for your ideas to get executed. Our product designers, engineering developers and innovative management teams ensure your product meets the world class standard. IP protection is at theheart of our management.

We follow a rigorous method to strictly protect your intellectual property rights in Asia and across the globeand offer you complete ownership of the design. We do not just stop by playing an advisory role. It is just our starting point. We walk along with you in executing these strategies end-to-end, to ensure business success.

Website: https://www.mitsquare.com

Industry: Information Technology & Services

Headquarters: London, UK

# CHAPTER-2

## EXECUTIVE SUMMARY
## DOMAIN NAME: EDU TECH

EduTech (a combination of "education" and "technology") refers to hardware and software designed to enhance teacher-led learning in classrooms and improve students' education outcomes. The goal of EduTech is to improve student outcomes, enhance individualized education, and reduce the teaching
burden on instructors. It encompasses several domains including learning theory, computer-based training, online learning and machine learning. EdTech is still in the early stages of its development, but it shows promise as a method of customizing a curriculum for a student's ability level by introducing and reinforcing new content at a pace the student can handle.

## 2.1 PROBLEM STATEMENT:

Here in our dataset, we are going to find the accuracy with the corresponding algorithm that works well for the given dataset and it also must be the best fit algorithm. In order to find the accuracy with best value, we may crossover several methods like (checking outliers, Data Pre-Processing etc…)

## 2.2 OVERVIEW OF THE PROJECT

## DATA DESCRIPTION:

This data set consists of the marks secured by the students in various subjects. To understand the influence of the parent's background, test preparation etc... on students' performance. This data set includes scores from three exams and a variety of personal, social, and economic factors that have interaction effects upon them. It helps to answer the questions such as,

- How effective is the test preparation course?
- Which major factors contribute to test outcomes?
- What would be the best way to improve student scores on each test?

## OBJECTIVE:

- Understand the Dataset & cleanup.
- Build Regression models to predict the students' marks.
- Evaluate the models & compare their accuracy scores.

## METHODOLOGY:

The different methods that we used to find the accuracy are given below

- ➢ **DATA VISUALIZATION**
- ➢ **LABEL ENCODING**
- ➢ **DATA PRE-PROCESSING**
- ➢ **PAIR PLOT**
- ➢ **TRAIN TEST SPLIT**
- ➢ **LINEAR REGRESSION**
- ➢ **SUPPORT VECTOR MACHINE**

## RESULT:

Maximum accuracy score from the dataset has been obtained by using support vector machine algorithm.

## 2.2.1 METHODOLOGY

## VISUALIZATION OF DATASET:

- **plot()**
  The plot() function is used to draw points (markers) in a diagram.
- **scatter**
  A scatter plot is a type of plot or mathematical diagram using Cartesian coordinates to display values for typically two variables for a set of data.
- **density**
  A density plot is used to visualize the distribution of a continuous numerical variable in a dataset. It is also known as Kernel Density Plots
- **histogram**
  A histogram is a graphical representation that organizes a group of data points into user-specified ranges**.**

## 2.2.2 LIBRARIES USED IN MACHINE LEARNING:

**NumPy**-which stands for Numerical Python, is a library consisting of multidimensional array objects and a collection of routines for processing those arrays. Using NumPy, mathematical and logical operations on arrays can be performed.

**Pandas**– It is a Python package providing fast, flexible, and expressive data structures designed to make working with "relational" or "labeled" data both easy and intuitive.Pandas is well suited for many different kinds of data: Tabular data with heterogeneously-typed columns, as in an SQL table or Excel spreadsheet.

**Matplotlib.pyplot**- It is a collection of functions that make matplotlib work like MATLAB. Each pyplot function makes some change to a figure: e.g., creates a figure, creates a plotting area in a figure, plots some lines in a plotting area, decorates the plot with labels, etc.

**Seaborn**-Seaborn is a data visualization library built on top of matplotlib and closely integrated with pandas data structures in Python. Visualization is the central part of Seaborn which helps in exploration and understanding of data.

# CHAPTER-3

## STUDENT'S MARKS ANALYSIS

## DATA SET:

In [2]:
```
import numpy as np
import pandas as pd
import seaborn as sb
import matplotlib.pyplot as mp



data=pd.read_csv("edu.csv")
data.head()
```

Out[2]:

| | gender | race/ethnicity | parental level of education | lunch | test preparation course | math score | reading score | writing score |
|---|---|---|---|---|---|---|---|---|
| 0 | female | group B | bachelor's degree | standard | none | 72 | 72 | 74 |
| 1 | female | group C | some college | standard | completed | 69 | 90 | 88 |
| 2 | female | group B | master's degree | standard | none | 90 | 95 | 93 |
| 3 | male | group A | associate's degree | free/reduced | none | 47 | 57 | 44 |
| 4 | male | group C | some college | standard | none | 76 | 78 | 75 |

In [43]:
```
data.shape
```

Out[43]: (1000, 8)

## ADDING EXTRA COLUMN TO THE DATASET:

**ADDING EXTRA COLUMN(AVERAGE)**

In [3]:
```
data['Average']=((data["math score"]+data["reading score"]+data["writing score"])/3).round()
data.head()
```

Out[3]:

| | gender | race/ethnicity | parental level of education | lunch | test preparation course | math score | reading score | writing score | Average |
|---|---|---|---|---|---|---|---|---|---|
| 0 | female | group B | bachelor's degree | standard | none | 72 | 72 | 74 | 73.0 |
| 1 | female | group C | some college | standard | completed | 69 | 90 | 88 | 82.0 |
| 2 | female | group B | master's degree | standard | none | 90 | 95 | 93 | 93.0 |
| 3 | male | group A | associate's degree | free/reduced | none | 47 | 57 | 44 | 49.0 |
| 4 | male | group C | some college | standard | none | 76 | 78 | 75 | 76.0 |

## 3.1 LABEL ENCODING:

Label Encoding refers to converting the labels into a numeric form so as to convert them into the machine-readable form. Machine learning algorithms can then decide in a better way how those labels must be operated. It is an important pre-processing step for the structured dataset in supervised learning.

**LABEL ENCODER**

In [4]:
```
from sklearn.preprocessing import LabelEncoder
l=LabelEncoder()
data["gender"]=l.fit_transform(data["gender"])
data["race/ethnicity"]=l.fit_transform(data["race/ethnicity"])
data["parental level of education"]=l.fit_transform(data["parental level of education"])
data["lunch"]=l.fit_transform(data["lunch"])
data["test preparation course"]=l.fit_transform(data["test preparation course"])

data.head()
```

Out[4]:

| | gender | race/ethnicity | parental level of education | lunch | test preparation course | math score | reading score | writing score | Average |
|---|---|---|---|---|---|---|---|---|---|
| **0** | 0 | 1 | 1 | 1 | 1 | 72 | 72 | 74 | 73.0 |
| **1** | 0 | 2 | 4 | 1 | 0 | 69 | 90 | 88 | 82.0 |
| **2** | 0 | 1 | 3 | 1 | 1 | 90 | 95 | 93 | 93.0 |
| **3** | 1 | 0 | 0 | 0 | 1 | 47 | 57 | 44 | 49.0 |
| **4** | 1 | 2 | 4 | 1 | 1 | 76 | 78 | 75 | 76.0 |

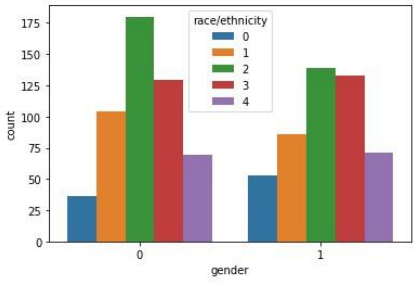
## 3.2 VISUALIZATION:

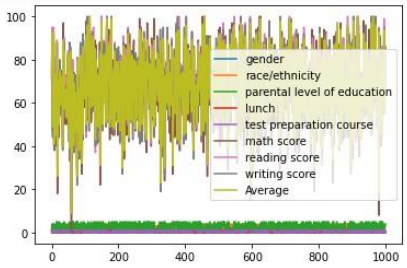
Data visualization is defined as a graphical representation that contains the information and the data. Data visualization provides an important suite of tools for identifying a qualitative understanding. This can be helpful when we try to explore the dataset and extract some information to know about a dataset and can help with identifying patterns, corrupt data, outliers, and much more.

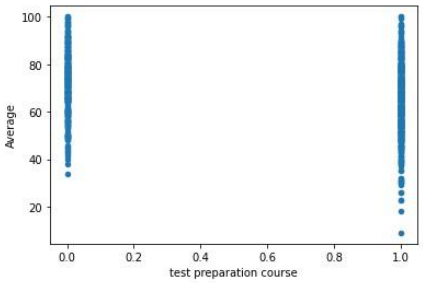
**VISUALIZATION**

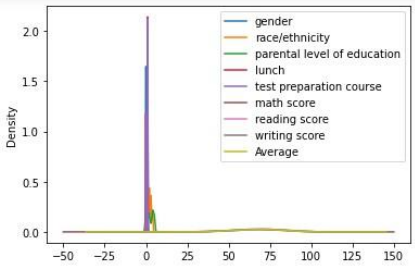
In [46]:
```
sb.countplot(data["gender"],hue=data["race/ethnicity"])
data.plot()
data.plot(kind="scatter",x="test preparation course",y="Average")
data.plot(kind="density")
data.plot(kind="hist")
```

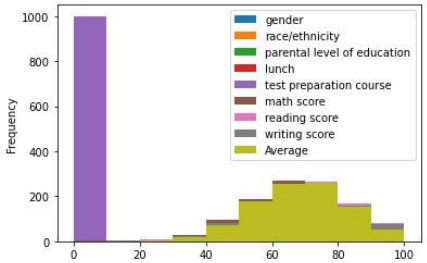
Out[46]: <AxesSubplot:ylabel='Frequency'>




gender
race/ethnicity
parental level of education
lunch
test preparation course
math score
reading score
writing score
Average


Average
test preparation course


Density
gender
race/ethnicity
parental level of education
lunch
test preparation course
math score
reading score
writing score
Average


Frequency
gender
race/ethnicity
parental level of education
lunch
test preparation course
math score
reading score
writing score
Average

## 3.3 DATA PRE-PROCESSING:

**DATA PRE-PROCESSING**

```
In [47]:
1 print(str("Any missing data or null in the dataset:"),data.isnull().values.any())
2 data.isnull().sum()

Any missing data or null in the dataset: False

Out[47]: gender                       0
race/ethnicity               0
parental level of education  0
lunch                        0
test preparation course      0
math score                   0
reading score                0
writing score                0
Average                      0
dtype: int64
```

**data.isnull()** in pandas will return True for missing components and False for non-missing cells.

```
In [5]:
1 data2=data.drop(columns=["lunch"])
2 data2.head()
```

Out[5]:

| | gender | race/ethnicity | parental level of education | test preparation course | math score | reading score | writing score | Average |
|---|---|---|---|---|---|---|---|---|
| **0** | 0 | 1 | 1 | 1 | 72 | 72 | 74 | 73.0 |
| **1** | 0 | 2 | 4 | 0 | 69 | 90 | 88 | 82.0 |
| **2** | 0 | 1 | 3 | 1 | 90 | 95 | 93 | 93.0 |
| **3** | 1 | 0 | 0 | 1 | 47 | 57 | 44 | 49.0 |
| **4** | 1 | 2 | 4 | 1 | 76 | 78 | 75 | 76.0 |

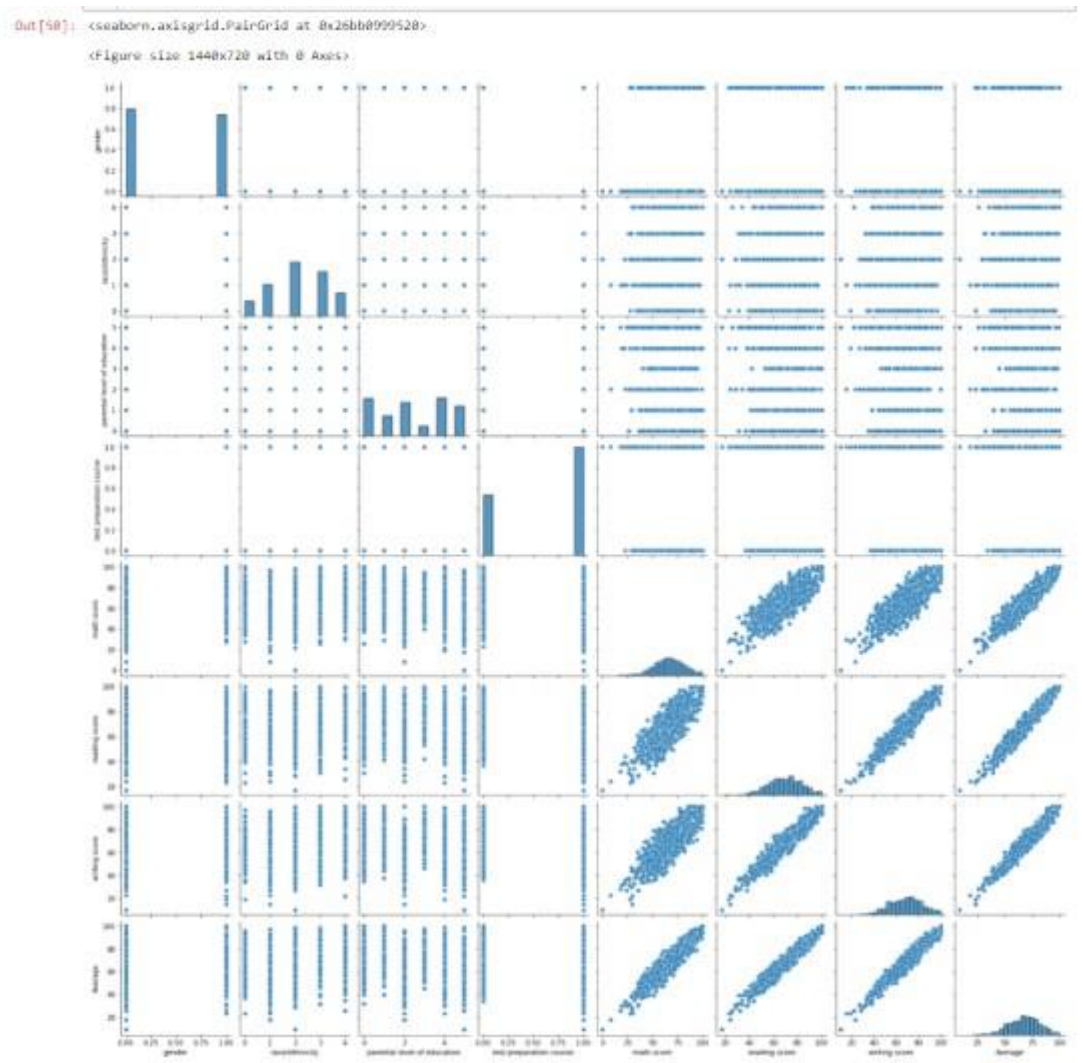
*[#We can drop the "lunch" column since it is not necessary]*

## 3.4 PAIR PLOT:

A pairplot plot a pairwise relationships in a dataset. The pairplot function creates a grid of Axes such that each variable in data will by shared in the y-axis across a single row and in the x-axis across a single column.

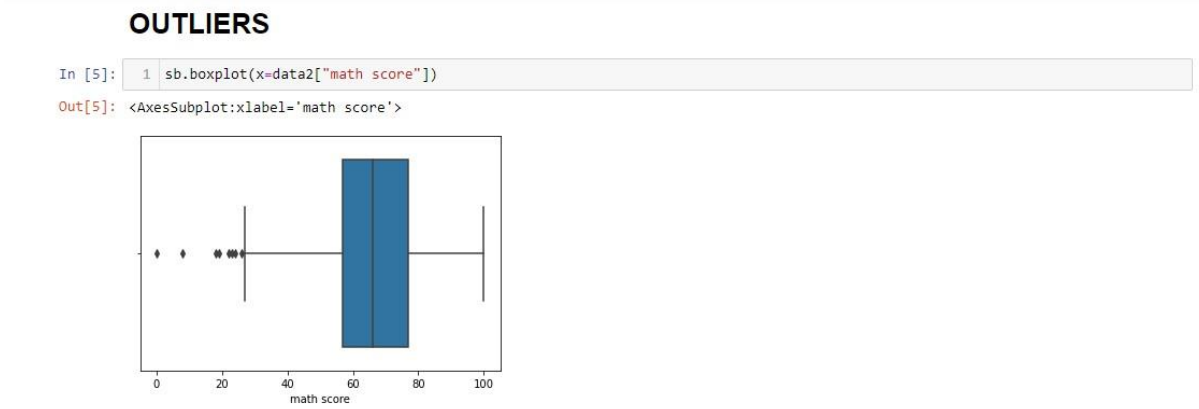
**PAIR PLOT**

```
In [50]:
1 mp.figure(figsize=(20,10))
2 sb.pairplot(data2)

Out[50]: <seaborn.axisgrid.PairGrid at 0x26bb0999520>

<Figure size 1440x720 with 0 Axes>
```




## 3.5 OUTLIERS:

An outlier is a data point that is noticeably different from the rest. They represent errors in measurement, bad data collection, or simply show variables not considered when collecting the data.



In [19]:
```
print(data2["math score"].quantile(0.10))
print(data2["math score"].quantile(0.90))
```

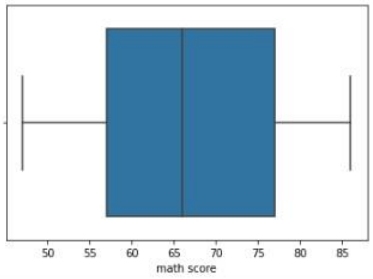
```
47.0
86.0
```

In [26]:
```
data2["math score"]=np.where(data2["math score"]<47.0,47.0,data2["math score"])
data2["math score"]=np.where(data2["math score"]>86.0,86.0,data2["math score"])
print(data2["math score"].skew())
```

```
0.004802837638950134
```

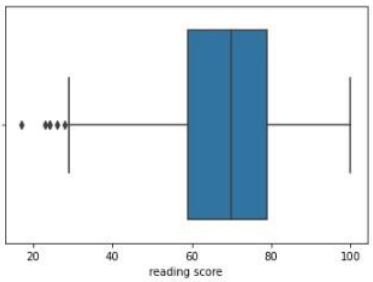
In [27]:
```
sb.boxplot(x=data2["math score"])
```

Out[27]: <AxesSubplot:xlabel='math score'>




In [7]:
```
sb.boxplot(x=data2["reading score"])
```

Out[7]: <AxesSubplot:xlabel='reading score'>




In [7]:
```
print(data2["reading score"].quantile(0.10))
print(data2["reading score"].quantile(0.90))
```

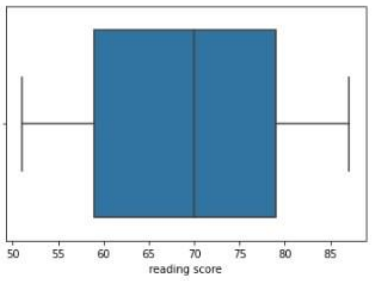
```
51.0
87.10000000000002
```

In [8]:
```
data2["reading score"]=np.where(data2["reading score"]<51.0,51.0,data2["reading score"])
data2["reading score"]=np.where(data2["reading score"]>87.10000000000002,87.10000000000002,data2["reading score"])
print(data2["reading score"].skew())
```

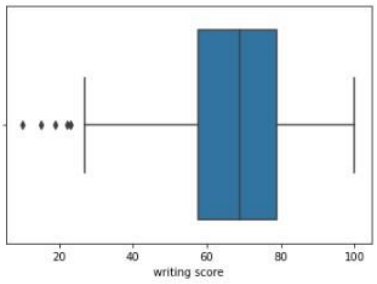
```
-0.056271207575291346
```

In [9]:
```
sb.boxplot(x=data2["reading score"])
```

Out[9]: <AxesSubplot:xlabel='reading score'>



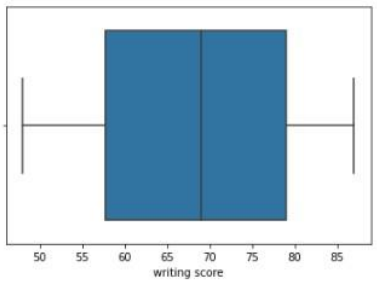
```
In [10]: 1 sb.boxplot(x=data2["writing score"])
Out[10]: <AxesSubplot:xlabel='writing score'>
```




```
In [11]: 1 print(data2["writing score"].quantile(0.10))
         2 print(data2["writing score"].quantile(0.90))

48.0
87.0
```

```
In [12]: 1 data2["writing score"]=np.where(data2["writing score"]<48.0,48.0,data2["writing score"])
         2 data2["writing score"]=np.where(data2["writing score"]>87.0,87.0,data2["writing score"])
         3 print(data2["writing score"].skew())

-0.11981484588886816
```

```
In [13]: 1 sb.boxplot(x=data2["writing score"])
Out[13]: <AxesSubplot:xlabel='writing score'>
```




## 3.6 TRAIN TEST SPLIT:

The train-test split is a technique for evaluating the performance of a machine learning algorithm. It can be used for classification or regression problems and can be used for any supervised learning algorithm. The procedure involves taking a dataset and dividing it into two subsets (Train and Test).

**TRAINING AND TESTING MODEL**

```
In [8]: 1 from sklearn.model_selection import train_test_split
        2 y=data2["Average"]
        3 x=data2.drop(["Average"],axis=1)
        4 x_train,x_test,y_train,y_test=train_test_split(x,y,test_size=0.2,random_state=0)
```

```
In [9]: 1 print("----------------TRAINING AND TESTING DATASETS----------------")
        2 print("x_train:",np.shape(x_train))
        3 print("y_train:",np.shape(y_train))
        4 print("x_test:",np.shape(x_test))
        5 print("y_test:",np.shape(y_test))

----------------TRAINING AND TESTING DATASETS----------------
x_train: (800, 7)
y_train: (800,)
x_test: (200, 7)
y_test: (200,)
```

*[We are going to use Regression algorithms as our dataset is not classifiable]*

**ALGORITHMS WE HAVE TAKEN**

➢ LINEAR REGRESSION

➢ SUPPORT VECTOR MACHINE

## 3.7 LINEAR REGRESSION

- Linear regression is a basic and commonly used type of predictive analysis. It is an approach for predicting a response using a single feature. It is assumed that the two variables are linearly related. Hence, we try to find a linear function that predicts the response value(y) as accurately as possible as a function of the feature or independent variable(x).
- In the simplest words, Linear Regression is the supervised Machine Learning model in which the model finds the best fit linear line between the independent and dependent variable i.e., it finds the linear relationship between the dependent and independent variable.

➢ **MSE** - It is the average of the squared error that is used as the loss function for least squares regression. It is the sum, over all the data points, of the square of the difference between the predicted and actual target variables, divided by the number of data points.

➢ **RMSE** – It is measured in the same units as the target variable. Due to its formulation, MSE, just like the squared loss function that it derives from, effectively penalizes larger errors more severely.

➢ **R2 score** - It is a very important metric that is used to evaluate the performance of a regression-based machine learning model. It is pronounced as R squared and is also known as the coefficient of determination. It works by measuring the amount of variance in the predictions explained by the dataset.

**LINEAR REGRESSION**

In [10]:

```
from sklearn.linear_model import LinearRegression
from sklearn import metrics
from sklearn.metrics import r2_score
from sklearn.model_selection import cross_val_score
from sklearn.metrics import mean_squared_error

model = LinearRegression()
model.fit(x_train,y_train)
accuracies=cross_val_score(estimator = model,X = x_train,y=y_train,cv=10)
y_pred = model.predict(x_test)

print('Score : %.4f' % model.score(x_test,y_test))
print('Mean score: %0.3f' % accuracies.mean())

mse = mean_squared_error(y_test,y_pred)
rmse = mean_squared_error(y_test,y_pred)**0.5
r2 = r2_score(y_test,y_pred)
print('------------------------------------------------------------')

print(' MSE : %0.2f ' % mse)
print(' RMSE : %0.2f ' % rmse)
print(' R2 : %0.2f ' % r2)
```

```
Score : 0.9996
Mean score: 1.000
------------------------------------------------------------
 MSE : 0.07
 RMSE : 0.27
 R2 : 1.00
```

## 3.8 SUPPORT VECTOR MACHINE (SVM REGRESSION)

- Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for Classification as well as Regression problems.

- The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes. so that we can easily put the new data point in the correct category in the future.

- SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called support vectors, and hence the algorithm is termed as Support Vector Machine.

**SUPPORT VECTOR MACHINE(SVM)**

In [12]:
```
from sklearn.preprocessing import StandardScaler
sc = StandardScaler()
x_train = sc.fit_transform(x_train)
x_test = sc.transform(x_test)

from sklearn.svm import SVR
from sklearn.preprocessing import StandardScaler
from sklearn.metrics import accuracy_score, confusion_matrix
from sklearn import preprocessing
import warnings

warnings.filterwarnings("ignore")

x_train, x_test, y_train, y_test = train_test_split (x,y,test_size=0.2)
svr = SVR()
accuracies = cross_val_score(svr,x_train,y_train,cv=20)
svr.fit(x_train,y_train)
#print(np.mean(accuracies))
print("ACCURACY:",svr.score(x_test,y_test))
```

```
ACCURACY: 0.9876080051098339
```

In [ ]:
```

```

## COMPARISON:

| LINEAR REGRESSION | SUPPORT VECTOR MACHINE |
|---|---|
| 1.Accuracy achieved in Linear Regression is 100% | 1.Accuracy achieved in Support Vector Machine is 98%. |
| 2.This model with 100% accuracy is said to be an overfitted model. | 2.This model with 98% accuracy is considered the best-fit model. |
| 3.As linear regression is found out to be over-fitted model, we neglect this accuracy score. | 3.Compared to linear regression, SVM gives best-fit accuracy. Hence this is our final accuracy. |

## 3.9 RESULT AND CONCLUSION

### RESULT:

- Support Vector Machine is the best-fit algorithm for our dataset as linear regression was failed to become a best-fit algorithm and being an over-fitted model.
- Accuracy achieved through SVM is **98%.**

### CONCLUSION:

- The size of the dataset we have taken was quite small with 1000 rows.
- To obtain more accuracy, we need to perform some operations such as Visualization, Removing outliers, etc...
- To find the best-fit accuracy score, we came across two algorithms (Linear Regression, Support Vector Machine).
- Linear Regression was not suitable for the dataset as it was over-fit model (100% accuracy).
- For our dataset Support Vector Machine gave the best accuracy score (98%) with the best-fit model

## REFERENCE:


- ✓ https://www.kaggle.com/spscientist/students-performance-in-exams
- ✓ https://www.analyticsvidhya.com/blog/2021/05/all-you-need-to-know-about-your-first-machine-learning-model-linear-regression/
- ✓ https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html
- ✓ https://www.geeksforgeeks.org/ml-label-encoding-of-datasets-in-python/
- ✓ https://seaborn.pydata.org/generated/seaborn.pairplot.html